Métricas de evaluación de modelos

Curso Aprendizaje Automático Aplicado 2026-1

Julio Waissman Vilanova

17 de febrero de 2026

¿Por qué evaluamos los modelos?

El objetivo de cualquier modelo de ML es la **generalización**.

Overfitting (Sobreajuste)
El modelo memoriza el ruido de los datos de entrenamiento y falla con datos nuevos.

Underfitting (Subajuste)
El modelo es demasiado simple para capturar la estructura subyacente.

División de los Datos (Split)

Para una validación robusta, dividimos el conjunto de datos en:

- **Training Set (Entrenamiento):** Usado para ajustar los parámetros.
- **Validation Set (Validación):** Usado para ajustar hiperparámetros y selección de modelo.
- **Test Set (Prueba):** Evaluación final e imparcial de la capacidad de generalización.

Error Cuadrático Medio (MSE) y RMSE

Miden la diferencia promedio entre los valores reales (y) y las predicciones ($\hat{y}$).

Fórmula del MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **RMSE (Root MSE):** Es la raíz cuadrada del MSE.
- **Propiedad:** Devuelve el error a las mismas unidades que la variable objetivo, facilitando la interpretación.

Error Absoluto Medio (MAE) y R-Cuadrado

MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Es menos sensible a valores atípicos (outliers) que el MSE.

Coeficiente de Determinación (R^2):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Representa la varianza explicada por el modelo (rango de $-\infty$ a 1).

El problema de clasificación

Se estima por ML:

$$\hat{a} = \mathsf{Pr}[y = 1|x; \theta]$$

Pero se quiere encontrar:

$$\hat{y} = \begin{cases} 1 & \text{si } a > u \\ -1 & \text{en otro caso} \end{cases}$$

Matriz de Confusión

Es la herramienta fundamental para analizar errores de clasificación.

	Predicho: Positivo	**Predicho: Negativo**
Real: Positivo	Verdadero Positivo (TP)	Falso Negativo (FN)
Real: Negativo	Falso Positivo (FP)	Verdadero Negativo (TN)

- **Exactitud (Accuracy):** $(TP + TN)$/Total. Engañosa en datasets desbalanceados.

Precisión y Exhaustividad (Recall)

Precisión (Precision)
¿Qué tan confiable es el modelo cuando dice que algo es positivo?

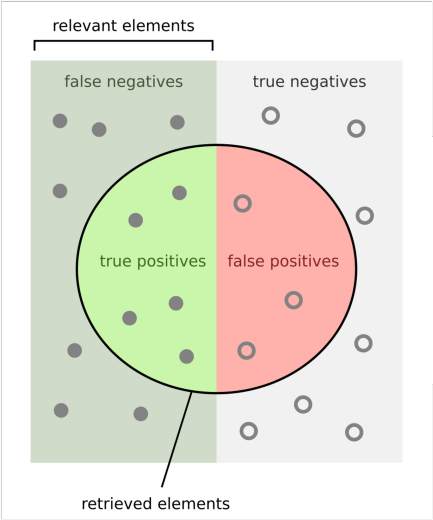
$$P = \frac{TP}{TP + FP}$$

Exhaustividad (Recall)
¿Qué porcentaje de los positivos reales detectó el modelo?

$$R = \frac{TP}{TP + FN}$$

F1-Score: Media armónica de ambas: $2 \cdot \frac{P \cdot R}{P + R}$

Todo en una imagen



How many retrieved items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

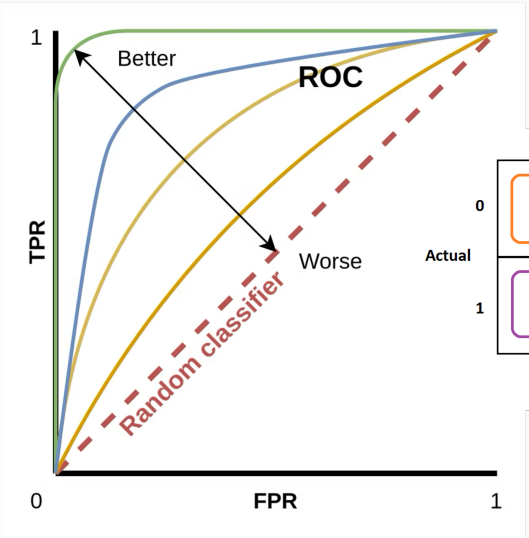
How many relevant items are retrieved?

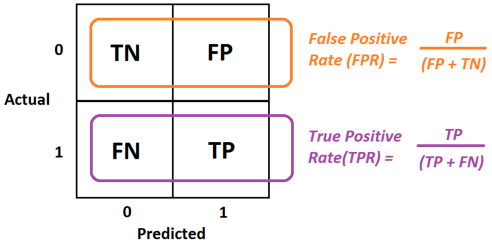
Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

Curva ROC y AUC

- **Curva ROC:** Grafica el Recall vs. Tasa de Falsos Positivos para distintos umbrales de decisión.
- **AUC (Area Under Curve):** Un valor de 1.0 indica un modelo perfecto; 0.5 indica un modelo que predice al azar.

Curva ROC y AUC





A tomar en cuenta

- **Regresión:** Usar MAE si hay muchos outliers; usar RMSE para penalizar errores grandes.
- **Clasificación Binaria:** No confiar solo en la Exactitud. Mirar F1-Score o AUC.
- **Problemas como Detección de Fraude/Cáncer:** Priorizar **Recall** (minimizar falsos negativos).
- **Problemas como Filtros de Spam:** Priorizar **Precisión** (minimizar falsos positivos en la bandeja de entrada).
- **Log-Loss:** Evalúa las probabilidades, no solo la clase final. Penaliza las predicciones seguras pero incorrectas.
- **Precision-Recall Curve:** Más informativa que ROC-AUC cuando hay clases muy desbalanceadas.

La Brecha de Comunicación

A menudo existe un divorcio entre lo que optimiza el científico de datos y lo que valora el ejecutivo.

- **Métricas de ML:** Definidas en función de la calidad del modelo *per se*.
- **Métricas de Negocio (KPIs):** Miden impacto (ROI, *Churn Rate*, Ahorro por Fraude, *Customer Lifetime Value*).

El Riesgo
Un modelo con un AUC de 0.95 puede ser inútil si no mejora el KPI de negocio o si su implementación es más cara que el beneficio que genera.

Frases Clave para la Armonización

- **”No optimices por precisión si el costo de un Falso Negativo es catastrófico”:** En diagnóstico médico, un 99% de precisión no sirve si el 1% de error son casos positivos omitidos.
- **”Habla el lenguaje del dinero, no el de las probabilidades”:** Traduce el incremento en el F1-Score a pesos ahorrados o ingresos adicionales generados.
- **”La métrica de ML es un medio, el KPI de negocio es el fin”:** Si la mejora técnica no mueve la aguja del negocio, el modelo requiere una reevaluación de objetivos.

Transformación: De Error Técnico a Costo de Negocio

Para armonizar, debemos asignar un **valor económico** a cada celda de la matriz de confusión.

Cálculo de Utilidad Esperada

$$U = (TP \times \text{Beneficio}) - (FP \times \text{Costo}) - (FN \times \text{Oportunidad Perdida})$$

- **Umbral de Decisión:** El umbral debe moverse hacia donde la *Utilidad Esperada* sea máxima para el negocio.
- **Aceptación:** Un modelo ”peor” técnicamente pero más interpretable puede ser preferible si reduce el riesgo legal o regulatorio.

Conclusión: El Círculo de Alineación

Diseño → Entrenamiento → Validación → Impacto

1. Definir el éxito del negocio primero.
2. Seleccionar la métrica de ML que más correlacione con ese éxito.
3. Validar no solo el error, sino el *Lift* (mejora respecto al proceso actual).
4. Monitorear en producción para asegurar que el valor se mantiene.